SPECTRAL AMPLITUDE WARPING (SAW) FOR NOISE SPECTRUM SHAPING IN AUDIO CODING

Roch Lefebvre, Claude Laflamme
University of Sherbrooke, Quebec, Canada, J1K 2R1
lefebvre@gel.usherb.ca

ABSTRACT

In this paper, we present a new approach to shape the coding noise in speech and audio coders. The approach, called Spectral Amplitude Warping (SAW), consists essentially of a pre- and post-processing which apply a non-linear transformation to the signal short-term spectrum prior to, and after, encoding. Since it is possible to view SAW as a separate entity from the coder, the noise shaping capability of an existing coder can be improved without modifying the coder itself. Using SAW as a pre- and post-process to the G.722 wideband speech coding standard, it was found in an informal listening test that the quality of the 64 kb/s operating mode can be achieved at only 48 kb/s. The price to be paid is an additional delay.

1. INTRODUCTION

It is widely recognized that to reduce the bit-rate of speech and audio coders without compromising quality, proper care has to be given to the spectral shape of the coding noise. A general rule is that the noise spectrum has to "follow" the signal spectrum [1]. This is known as noise shaping.

In low bit-rate speech coders (< 16 kb/s), which typically use the CELP model [2], a time-varying perceptual filter controls the noise level as a function of frequency. This perceptual filter is derived from an all-pole filter which models the formants, or spectral envelope, of the speech spectrum. Hence, the noise spectrum approximately follows the speech formants. Further perceptual improvements can be achieved by using a post-filter [3], which emphasizes the formant and harmonic structure of the synthesized speech signal.

In higher bit-rate audio coders, which are typically transform or sub-band coders, the noise spectrum is controlled by dynamic bit allocation in the frequency domain. In the most complex algorithms, a sophisticated hearing model is used to determine a masking threshold. The bit allocation is such that the distortion remains below the masking threshold at any frequency [1] (provided the coder operates at a sufficient bit-rate). The resulting coding noise will be correlated to the signal spectrum, with corresponding peaks and valleys.

Spectral Amplitude Warping (SAW), the noise shaping algorithm presented in this paper, is based on this observation that the perceptually optimal noise spectrum should be highly correlated to the signal spectrum. SAW can be viewed as a pre- and post-process which apply a non-linear transformation to the signal spectrum prior to (and after) encoding. The non-linear transformation has to be chosen so that the coding noise is properly shaped.

Section 2 first reviews the basic idea of coding with pre-emphasis. Then, Section 3 presents the principle of SAW. An application of SAW is presented in Section 4, where it is shown how it can improve the noise shaping capability of an existing coder, namely the G.722 wideband speech coding standard. Listening tests results are presented in Section 5. Finally, Section 6 gives some conclusions, along with other possible uses of the SAW algorithm.

2. CODING WITH PRE-EMPHASIS

Consider Figure 1, where the encoding of signal $s(n)$ is split into three basic functions. First, a pre-emphasis function which transforms input signal $s(n)$ into signal $s_w(n)$. Then, an encoding (quantization) operation, represented here by a simple addition where $q(n)$ is the coding noise in the pre-emphasized domain. Finally, a de-emphasis function which performs the inverse of the pre-emphasis function.

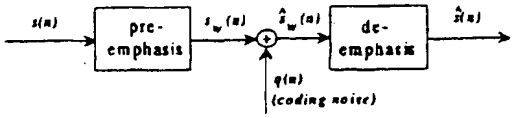

Figure 1. Coder with pre-emphasis.

If the pre-emphasis and de-emphasis functions are linear filters whose transfer functions are $W(z)$ and $W(z)^{-1}$, respectively, then it can be shown that

$$\hat{S}(z) = S(z) + Q(z)\,W(z)^{-1} \qquad (1)$$

where $Q(z)$ is the z-transform of the coding noise $q(n)$. Equation (1) implies that the noise spectrum between $s(n)$ and $\hat{s}(n)$ is given by $Q(z)\ W(z)^{-1}$, evaluated at $z = e^{j\omega}$. Essentially, the coding noise $Q(z)$ is shaped by filter $W(z)^{-1}$.

Because of the non stationarities of speech and audio, $W(z)$ must be adaptive to follow the signal statistics. This adaptation implies either that forward information is sent to the decoder (to adapt the deemphasis filter $W(z)^{-1}$

0-8186-7919-0/97 $10.00 © 1997 IEEE

accordingly), or that the adaptation is based on the past synthesis signal (backward adaptation).

3. PRINCIPLE OF SPECTRAL AMPLITUDE WARPING

Consider then the following transformation as pre-emphasis in Figure 1:

$$S_w(k) = f_{nl}(S(k)) \qquad (2)$$

where $S(k)$ and $S_w(k)$ are the Fourier spectra (or any other time-frequency mapping such as the DCT) of signals $s(n)$ and $s_w(n)$, respectively, and where $f_{nl}(S(k))$ is a non-linear function of $S(k)$. In this paper, we investigate the following non-linear transformation:

$$S_w(k) = S(k) \frac{|S(k)|^{\alpha(k)}}{|S(k)|} \qquad (3)$$

where $\alpha(k)$ is a constant, possibly the same for all k. If for example $\alpha(k)=\alpha=0.5$, then $|S_w(k)|$ will be the square root of $|S(k)|$, with the phases unchanged. The effect of this transformation is to make $|S_w(k)|$ more flat than $|S(k)|$, as the attenuation factor for large amplitudes will be greater than for small amplitudes. If the resolution of $S(k)$ is sufficient, this effect will be two-fold:

(1) attenuation of the formants;
(2) attenuation of the harmonics.

In other words, the differences between peaks and valleys will be attenuated at both the "macroscopic" (formant) and "microscopic" (harmonic) levels.

Referring again to Figure 1, signal $s_w(n)$ is then corrupted with coding noise $q(n)$, and the synthesis signal $\hat{s}(n)$ is obtained by applying the inverse non-linear transformation to $\hat{S}_w(k)$:

$$\hat{S}(k) = \hat{S}_w(k) \frac{|\hat{S}_w(k)|^{(1/\alpha(k))}}{|\hat{S}_w(k)|} \qquad (4)$$

Since $\hat{S}_w(k) = S_w(k) + Q(k)$, it is easy to show that $\hat{S}(k) = S(k)$ when $Q(k) = 0$ by simply letting $\hat{S}_w(k) = S_w(k)$ in Equation 4.

Further, using Equations 3 and 4 we can write the following relation between $S(k)$ and $\hat{S}(k)$:

$$|\hat{S}(k)|^{(\alpha-1)}\hat{S}(k) = |S(k)|^{(\alpha-1)}S(k) + Q(k) \qquad (5)$$

The term on the left-hand side is simply $\hat{S}_w(k)$, and the first term on the right-hand side is $S_w(k)$.

It is evident that in Equation 5, $\hat{S}(k)$ can not be expressed explicitely as a function of $S(k)$, as in Equation 1. However, if we make the assumption that the noise $Q(k)$ is sufficiently smaller than the pre-emphasized signal $S_w(k)$, then we have the approximation

$$|\hat{S}(k)|^{(\alpha-1)} \approx |S(k)|^{(\alpha-1)}$$

and Equation 5 becomes

$$\hat{S}(k) \approx S(k) + Q(k)|S(k)|^{(1-\alpha)} \qquad (6)$$

which means that the noise spectrum between $s(n)$ and $\hat{s}(n)$ is now shaped by $|S(k)|^{(1-\alpha)}$, which is obviously correlated to the spectrum of the input signal. For example, if $\alpha=0.5$, then the noise spectrum will be shaped by the square-root of the input signal amplitude spectrum. This factor $|S(k)|^{(1-\alpha)}$ in Equation 6 plays the same role as filter $W(z)^{-1}$ in Equation 1.

So far, we have looked at SAW in the frequency domain, to get an understanding of how noise is shaped. However, in the general framework of Figure 1, the pre-emphasis and de-emphasis functions should produce time-domain signals with the proper spectral caracteristics. This can be accomplished in a number of ways. Figure 2 shows the overlap-and-add approach.

The original time-domain signal $s(n)$ is first windowed to minimize the effect of discontinuities. The frequency resolution of the noise shaping in SAW is a direct function of this window size. Then the spectrum $S(k)$ is computed using an N-point FFT (if the Fourier transform is used). $S_w(k)$ is then obtained as in Equation (3); this is referred to as non-linear warping in Figure 2. The inverse FFT gives the time-domain representation of $S_w(k)$. To allow a smooth transition between two successive blocks of signal $s_w(n)$, we use overlap-and-add with a hanning window (window 2) to reconstruct $s_w(n)$. This implies that window 1 has a flat portion equal at least to the length of window 2. If window 2 is centered at the middle of window 1, then the delay of SAW will be exactly half the length of window 1.

To perform the de-emphasis function of Figure 1, Equation 4 is used as the non-linear warping in Figure 2. In this case, the input and output are $\hat{s}_w(n)$ and $\hat{s}(n)$.

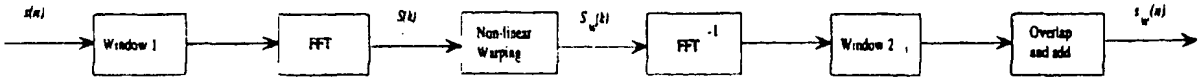

Figure 2. Principle of Spectral Amplitude Warping (SAW) with overlap-and-add.

4. APPLICATION TO G.722

To illustrate the noise shaping capability of SAW, we used it as a pre- and post-process to the G.722 wideband speech coding standard [4]. The G.722 coding algorithm is essentially a two sub-band coder with ADPCM encoding in each sub-band. Each sub-band is decimated by a factor of 2 prior to encoding. The high-frequency band (4-8 kHz) is coded with 2 bits/sample while the low-frequency band (0-4 kHz) is coded with 4, 5 or 6 bits/sample. The total bit-rate is thus 48, 56 or 64 kb/s. The coding noise in each sub-band is approximately white.

Figure 3 shows the amplitude spectrum of a voiced speech segment. Figure 4 shows the coding noise for the same segment, using G.722 at 48 kb/s. It can be seen that the coding noise in each band is approximately flat. Further, in some regions of the spectrum, the coding noise exceeds the original signal spectrum, in particular between 2 and 5 kHz. This results in audible distortion, as confirmed by listening to the synthesized speech.

Figure 5 shows the resulting coding noise if we use G.722 at 48 kb/s to code signal $s_w(n)$ in Figure 1, with the pre-emphasis function given by Equation 3 and the de-emphasis function given by Equation 4, and $\alpha(k) = \alpha = 0.5$. A 512-point FFT is used, resulting in a 16 ms delay at the coder, and 16 ms delay at the decoder. The length of window 2 in Figure 2 is 128 samples.

Note that the noise spectrum in Figure 5 is highly correlated to the signal spectrum in Figure 3, with corresponding formants and harmonics. In this case, the synthesis is perceptually very close to the original, with almost no audible artifacts.

It is worthwhile noting that even though SAW noise shaping is signal adaptive, the pre-emphasis and de-emphasis functions are constant functions which do not themselves require adaptation. Hence, there is no additional side information to be sent as in forward adaptation, and the local synthesis need not be known at the coder as in backward adaptation. The SAW pre-emphasis and de-emphasis functions can be completely separated from the coder. This is a definite advantage in attempting to improve the noise shaping capability of an existing coder without even modifying the coder itself.

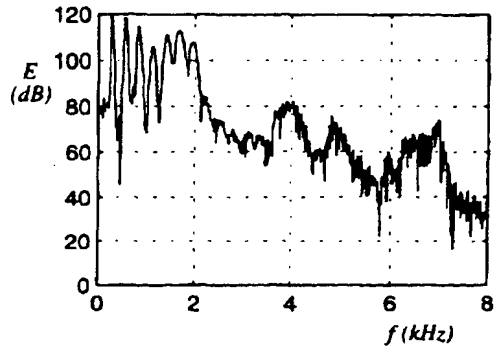

Figure 3. Original speech spectrum.

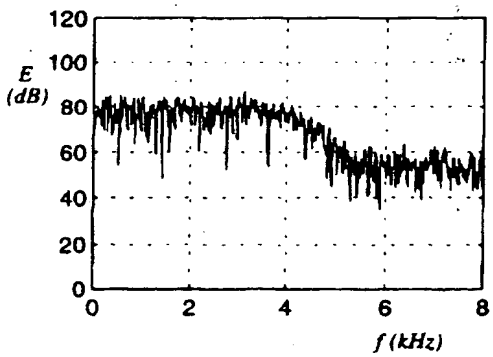

Figure 4. Spectrum of error signal for G.722 at 48 kb/s.

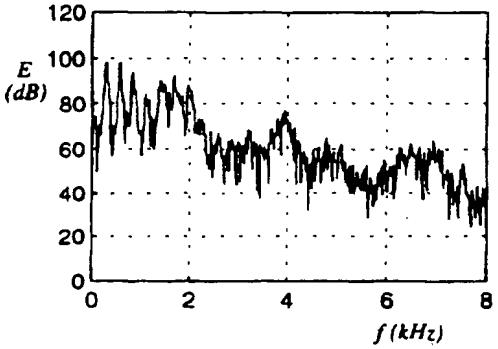

Figure 5. Spectrum of error signal for G.722 at 48 kb/s with SAW.

5. LISTENING TESTS RESULTS

Informal listening tests were conducted to assess the quality improvement obtained by applying SAW to the G.722 wideband speech coding standard. The test involved ten listeners. Subjects were presented a series of speech or music signal pairs. The test files consisted of 4 speech files and 4 music files. In each pair, one of the signals corresponded to the G.722 coder at 48 kb/s with SAW (the value α=0.6 was chosen as an optimal compromise for all types of files), and the other signal corresponded to the G.722 coder at either 48 kb/s or 64 kb/s, without SAW. Hence, there was always one signal in a pair which had been processed through SAW. The order in which this file appeared to the listener was random from one pair to the next.

For each signal pair, listeners were asked to indicate which signal had the best quality. A possible answer was also that both signals had the same quality. The test was conducted using headphones. Listeners were allowed to listen to each signal as often as they wanted before indicating their preference.

The results of the listening tests are shown in Table 1. "Enhanced G.722" refers to the G.722 coder with SAW used as pre- and post processing. "Reference coder" refers to G.722 without SAW.

The first line of Table 1 compares the quality of G.722 at 48 kb/s with and without SAW. Results show a dramatic improvement in quality when using SAW (79% preference rate). Note that the files for which the quality was judged to be the same are mostly the castanet and pop music sequences. For most speech files, as well as classical and solo instruments, listeners preferred the enhanced coder.

The second line of Table 1 gives a better indication of the increase in quality obtained with SAW. Here, we see that 83% of the time, the enhanced G.722 at 48 kb/s is judged to be of equal or better quality than the G.722 coder at 64 kb/s. Thus, for most signals tested, using SAW as a pre- and post-process to the G.722 coder makes it possible to achieve the quality of the 64 kb/s operating mode at only 48 kb/s.

Reference coder	Prefers reference	Same	Prefers enhanced G.722 at 48 kb/s
G.722 at 48 kb/s	0 %	21 %	79 %
G.722 at 64 kb/s	17 %	54 %	29 %

Table 1. Listening tests results.

6. CONCLUSION

A new noise shaping approach for speech and audio coders was presented in this paper. The approach, called Spectral Amplitude Warping (SAW), is implemented as a pre-emphasis and de-emphasis operation which modify the signal spectrum prior to (and after) encoding. The fonction used to modify the spectrum is a non-linear transformation, so that a constant fonction, which does not require additional bits to be sent, can achieve signal adaptive noise shaping.

There are many advantages to this approach. One advantage, demonstrated in Sections 4 and 5, is that the noise shaping capability of a coder can be improved without even modifying the coder itself. Even coders which do not have noise shaping capabilities can be improved upon when used in conjonction with SAW. For example, experiments were conducted using a simple μ-law PCM coder and SAW. Nearly transparent quality can be obtained at rates as low as 4 bits/sample for 48 kHz audio (192 kb/s). Since the pre-emphasized signal in SAW still retains much of its original redundancy (harmonics and formants will still be present, but to a lower extent), a more intelligent coder could sustain this quality at much lower rates. Other experiments involved low bit-rate CELP speech coders, and showed that the quality for music, in particular, can be significantly improved provided the necessary delay for SAW can be accomodated.

Finally, the SAW approach provides a framework where noise shaping and coding per se are separate operations. This makes it possible to use a coder based on short frames (to optimize pitch prediction, for instance), while using a higher frequency resolution (longer frames) to perform noise shaping.

7. REFERENCES

[1] N.S. Jayant, J.D. Johnston, Y. Shoham, "Coding of Wideband Speech," *Speech Communication*, Vol. 11, No2. 2-3, June 1992.

[2] M.R. Schroeder, B. Atal, "Code-Excited Linear Prediction (CELP): High-quality speech coding at very low bit rates," *Proc. IEEE ASSP*, pp. 937-940, 1985.

[3] A. Gersho, "Advances in Speech and Audio Coding," *Proc. IEEE*, Vol. 82, No, 6, June 1994.

[4] X. Maitre, "7 kHz Audio Coding within 64 kbit/s," *IEEE Jour. Sel. Areas in Comm.*, Vol 6, No. 2, February 1988.